NOAA Technical Memorandum NWS SR-220

A CATEGORICAL FLOOD FORECAST VERIFICATION SYSTEM FOR SOUTHERN REGION RFC RIVER FORECASTS

Robert J. Corby
West Gulf River Forecast Center
Fort Worth, Texas

William E. Lawrence
Arkansas-Red Basin River Forecast Center
Tulsa, Oklahoma

Scientific Services Division
Southern Region
Fort Worth, Texas

June 2002

UNITED STATES
DEPARTMENT OF COMMERCE
Donald L. Evans, Secretary

National Oceanic and Atmospheric Administration
Conrad C. Lautenbacher, Jr.
Under Secretary and Administrator

National Weather Service
John J. Kelly, Jr.
Assistant Administrator
for Weather Services

1. Introduction

The National Weather Service (NWS) River Forecast Centers(RFC) are responsible for producing flood forecasts for major rivers and streams throughout the country. For some time now, there has been a recognized need for a river forecast verification system to evaluate the RFC's skill in the delivery of this service. The agency cites two goals for verification: (1) Improve accuracy and timeliness of river forecasts and (2) document overall trends in forecast performance (NWS, 2001). Verification of river forecasts at individual locations are required to identify areas where forecast skill requires improvement. Hydrologic development efforts can then be concentrated on areas where an RFC consistently does an inadequate job of forecasting. Accumulating pertinent verification information for an RFC over a period of time, an RFC can show trends in their overall river prediction performance. As technology and science continue to advance, "improvements" will continue to be made to our hydrologic models and their underlying infrastructure. We are now capable of generating river forecasts quicker and at higher resolution than ever before. Without concrete performance measures, it is difficult to link improvements in river forecast accuracy to technological and scientific changes.

With these two goals in mind, the Southern Region Categorical Flood Forecast Verification System(CFFVS) was developed. The system is designed to provide a meaningful analysis of flood forecast performance at the individual forecast point scale, as well as over a larger scale, such as an RFC service area or a regional service area.

2. Background

There has never been a national RFC river forecast verification system in place until the recent initiative by the Office of Services, Hydrologic Services Division (OS/HSD), which was influenced by the agency's increased emphasis on performance measures. Verification methodology has been discussed over the years, but chiefly due to the complexity of the problem, a consensus has never been reached on the proper metrics necessary to measure forecast performance.

Several river forecast verification initiatives have been implemented over the years. Most of these schemes, including the current NWS national verification software, measure river forecast skill based on statistical error and bias. Commonly expressed in terms of absolute error or root mean square error (RMSE), these methods measure the difference between forecasted and observed stages. These types of statistics serve a useful purpose of comparing forecasts at a single location, or tracking model performance. For instance, the Arkansas-Basin River Forecast Center (ABRFC) used RMSE to evaluate model performance with and without QPF (Reed et al.1997). This information can be valuable when comparing different forecast techniques, but is generally meaningless when aggregated for multiple sites. RMSE is small during recession and baseflow conditions, but increases greatly during river rises. The ABRFC has shown that over a three-year period, there was a .91 correlation between monthly flows and forecast error (Wasko et al. 2001). Basically, RMSE applied to river forecast verification simply shows when it rained, and when it didn't. Clearly, forecast error statistics alone is no way to convey how well an RFC did in providing routine river forecast services.

In 1988, the original concept of a categorical flood forecast verification system was presented in the NOAA Technical Memorandum, NWS HYDRO-43 (Morris,1988). Morris suggested that floods are events, similar to weather phenomena such as tornadoes and hurricanes. As events, floods can be classified as to magnitude: minor, moderate, or major. With categories defined, we can readily verify the magnitude of a flood, and verify how accurate our forecasts were with respect to these established categories. For instance, if we observed a major flood, did we forecast a major flood? If so, how much lead time did the forecast provide? If we didn't, how badly did we miss forecasting a major flood? By answering these basic questions we "frame the service," or provide statistics relative to the hydrologic significance of the flood event.

In December, 1999 under the direction of Southern Region Hydrologic Services, a verification team was formed with members from ABRFC, WGRFC, and SRH. The team was tasked with the implementation of a categorical flood verification scheme based on the concepts contained in NOAA TM NWS Hydro-43. Additional requirements were placed on the team to keep the process "simple and automated." After several iterations, the verification team agreed on a proposed set of verification metrics in late 2000. In January 2001 the team was expanded to include a representative from the remaining SR RFCs and the senior service hydrologist from the WFO in Norman, Oklahoma.

The software was installed at all SR RFCs in June 2001, and currently verification statistics are computed quarterly at each RFC and consolidated on a regional Web site.

3. Categorical Flood Forecast Verification System

CFFVS is set up to run as a local application on the Advanced Weather Information Processing System (AWIPS) at an RFC. The verification process can be broken down into three discrete steps:

1. Data Assimilation
2. Forecast Verification
3. Executive Summaries

Data assimilation begins with compilation of category thresholds to use for the verification program. It also includes the ongoing archival of forecast and observed river stages. The forecast verification step consists of the computation and output of categorical statistics for each individual forecast point. The final step produces executive summaries which are computed from overall statistical scores for a designated time period. These aggregated scores, can then be compiled and presented to management. Each of these steps is discussed in detail below.

3.1 Categorical Data Assimilation

Before implementation of a categorical verification scheme, the category levels for each forecast point must be determined. The CFFVS uses the flood severity categories of minor, moderate, and major to classify flood forecasts. We have further stratified forecast locations, by adding an action stage category, since many RFCs use this stage as a threshold for initiating flood forecasts, and a record flood category, to provide specific verification information for the most extreme events. Table 1 provides the flood category information used in the CFFVS and the RFC AWIPS database files that contain this information.

Category	Severity	Reference Source
0	Action Stage	OFS Rating Curve File
1	Minor Flooding (Flood Stage)	OFS Rating Curve File
2	Moderate Flooding	IHFS Floodcat Table
3	Major Flooding	IHFS Floodcat Table
4	Record Flooding	OFS Rating Curve File

Table 1. Flood Categories.

In the Southern Region, the service hydrologists set action stages and flood stages as well as the threshold stages for minor, moderate, and major flooding. There are Integrated Hydrologic Forecast System (IHFS) database fields available in the WFO hydrologic database for each of these stages. The service hydrologists also are responsible for determining the flood of record for all locations. Each service hydrologist tabulates this pertinent data for all forecast points in their Hydrologic Service Area(s) (HSA) and provides it to the appropriate RFC. The RFC then enters the data into their IHFS database and OFS rating curve files. Once the categorical information from the WFOs is assimilated at the RFC, it is envisioned that only infrequent updates will be required for the categorical information.

Although desirable, it is not required to have valid entries for each category. CFFVS will stratify forecasts into any and all categories that are defined. For instance, if no action stage is defined in the OFS rating curve file, the lowest verification category will be minor flooding. There may be instances where the flood of record is lower than some of the other categorical stages. This is most likely to occur at forecast points that have a short historical record. In this case, the record flood category is ignored. Record flood level must exceed the major flood category to be considered a valid flood category.

3.2 Archive Database

The second, and ongoing data assimilation requirement, is archiving the observed and forecast data for CFFVS.

Both observed and forecast stage data are processed from products coded in the Standard Hydrometeorological Exchange Format(SHEF) (NWS,1998). Currently, all SR RFCs prepare their forecasts using the XSETS software. XSETS, which stands for Xwindows based SHEF Encoded Time-Series, was originally developed at ABRFC, and is now supported nationally by the Office of Hydrologic Development (OHD). This software extracts forecast time-series generated by the NWS River Forecast System (NWSRFS), and reformats them into SHEF. An example of an XSETS forecast is shown in Appendix A. The SHEF products are disseminated to Weather Service Forecast Offices (WSFOs) where they are used as guidance in the preparation of public river forecasts. The products are also passed to a SHEF decoder at the local RFC where they are stored in the IHFS database. Each forecast ordinate is stored in a separate database record in the IHFS*fcstheight* table. Each record in fcstheight contains a gauge height value, a *valid time* which is the date and time for the forecast ordinate, and a *basis time*, which is the date and time the forecast was issued.

Observed gauge height data are also ingested into the IHFS database. Products containing SHEF encoded observations are processed through the SHEF decoder and posted to the *height* table in the IHFS database. As with the forecast data, each individual observation is stored as a record in the database.

Before we can verify a forecast, both the forecast and observed data must be moved from the operational IHFS database to the archive database. The archive database structure we have implemented simply mimics the database structure of the fcstheight and height tables in the IHFS database. Appendix B shows the database schema for the archive data. Standard query language (SQL) procedures are used to copy data to the archive database. In our implementation, we initiate the SQL script daily to populate the archive database. The SQL script archives all forecast time series, and all hourly gauge height observations.

Even though we archive all hourly observations, the CFFVS examines only the observations that have identical times as the forecast ordinates. Since most of our forecasts produce six hour ordinates at synoptic times, six hour observations at 0000,0600,1200, and 1800 UTC would suffice in most cases. The intermediate observations may be useful to estimate missing synoptic time data, but are not required if disk space is limited. It should be noted that if an observation does not exist in the archive database at the same time of a forecast ordinate, no verification is attempted. Also, CFFVS does not provide any data quality control. It is up to the RFC to ensure that bad observations have been removed, and any interpolated stages have been entered. ABRFC has developed a couple of utility programs to aid in data inspection, but currently there are no user applications to correct erroneous data in the archive database.

4. Forecast Verification

The CFFVS is a derivative of the methods outlined in NOAA TM NWS Hydro-43. In order to make the process "automated and simple" the methods were adapted to utilize time series forecasts and observations. When the original concept was introduced in 1988, forecast and observed data were not readily available in a format which could quickly and easily be processed by automated methods. RFCs prepared river forecasts in a text format, using qualitative terms to describe the shape and magnitude of the forecast hydrograph. Commonly the forecasts contained statements such as "crest near," "rise to," or were bracketed between an upper and lower stage using terminology like "crest between." Timing of the crest was also generalized, using broad terms such as "midday" or "late evening," or even "next Tues." Other terms were included in the forecast to describe the shape of the hydrograph. These were usually expressed as times to rise above or fall below critical stages, such as bankfull or flood stage. While these statements legitimately expressed the uncertainty in the flood prediction process, they made verification subject to interpretation of what the forecaster was trying to convey. The sparsity of observed data, also made verification difficult to implement. While real-time data at frequent intervals was quickly becoming available, daily gauge height readings from COOP observers provided the only observations at many locations.

4.1 Verification Definitions

Before explaining the methodology behind CFFVS, it is important to understand the following definitions:

Forecast: Each forecast ordinate of each forecast time series is verified as its own forecast. If a three-day forecast is issued with six hour ordinates, 12 separate forecasts will be evaluated. If the forecast is updated six hours later, another 12 forecasts will be verified. Any reference to a forecast in these definitions should be construed as an individual forecast ordinate.

Flood Category: Each individual forecast point is stratified into flood categories ranging form action stage to flood of record. Forecasts below action stage are not verified. Category stages must increase as categories progress from action stage to flood of record.

Hit: If a forecast and corresponding observation at the same valid time are in the same flood category, the forecast is credited with a hit.

Lead Time: A categorical lead time is the number of hours from the time of forecast issuance to the time of the forecast hit. A lead time is only computed when, (1) the ordinate's forecast and observation are in the same category, and (2) the previous ordinate's observation was lower than the current category. This restricts lead time calculations to instances where the stage is rising, and crossing categories.

Miss: A miss is given when a forecast is in one category, and the corresponding observation at the same valid time is in a different category.

Categorical Error: This is the amount the forecast would have to be changed to reach the observed category. Categorical error is only computed when you have a miss.

False Alarm: A false alarm is given when the forecast is above the lowest defined category, and the observation is below the lowest category.

No Forecast Miss: A no forecast miss is the least desirable score. In this case, the observation is above the lowest defined category, and there is no corresponding forecast ordinate. In this case flooding occurs before a forecast is issued.

Non-Flood Forecast: Any forecast that is below the lowest defined category with an observed stage also below the lowest defined stage is a non-flood forecast. Non-flood forecasts are not counted in the verification process.

Number of Events: The number of events is determined by summing the total number of hits, misses, false alarms, and no forecast misses for a specified period.

4.2 Verification Example

Multiple forecasts are often issued for a particular flood event (Fig. 3). These forecasts may be updated every six to 12 hours, depending on policy guidelines and how well the forecast is tracking. Figure 1 depicts a case where three time series forecasts were issued for a hypothetical event. Each forecast time series is verified with observed data. For our example we will examine the time series forecast labeled Fcst 1 (Fig. 2). Each forecast ordinate for Fcst 1 is evaluated against the observed stage at the same valid time. Each forecast ordinate is classified as a hit, miss, no forecast miss, or false alarm within the pertinent category. Lead times are calculated for hits where the previous observation was below the current category, and categorical error is calculated for misses (Table 2). The verification procedure is repeated for Fcst2 and Fcst3, and the results are tallied for each category. Figure 4 contains a generalized process flow.

Much thought was put into the lead time computation. After several iterations, we arrived at the current method which seems to provide a meaningful value of lead time. The requirement that the previous ordinate's observation fall in a lower category isolates the cases where the river has changed categories, and is on the rising limb. This limits the number of cases where lead time is calculated, but provides an accurate depiction of lead time within a category.

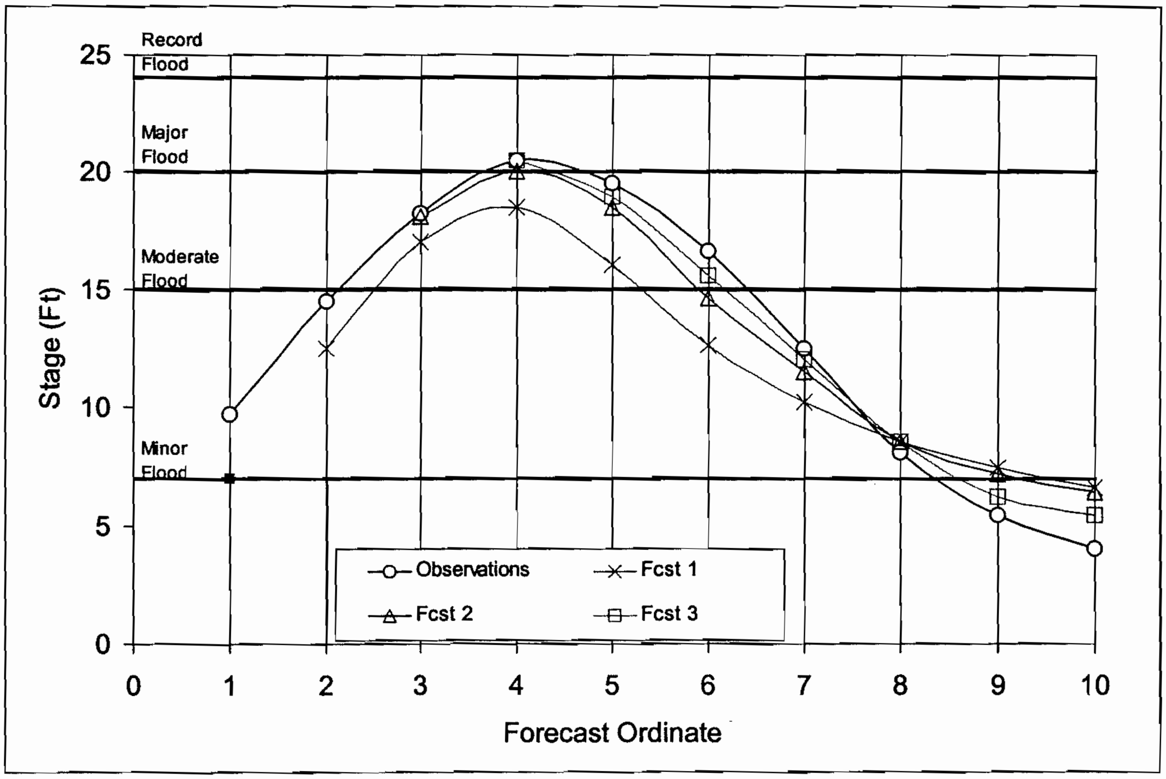


Figure 1 Example of multiple forecast issuances.

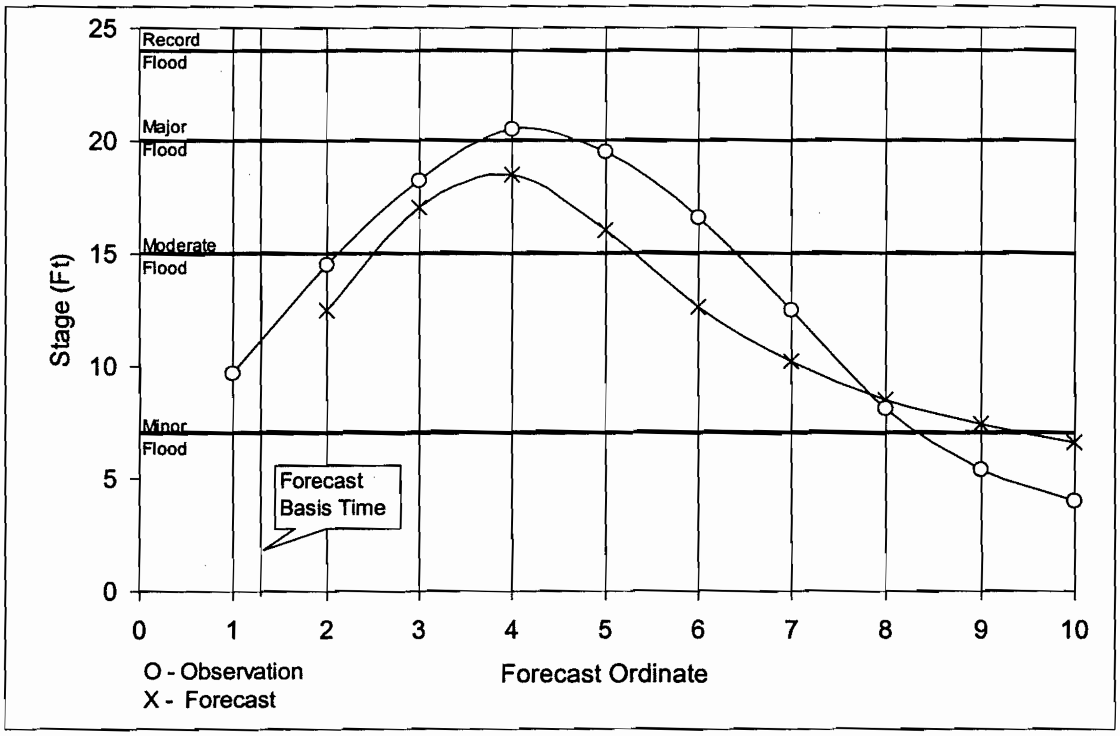


Figure 2 Example of a single forecast issuance (Fcst 1).

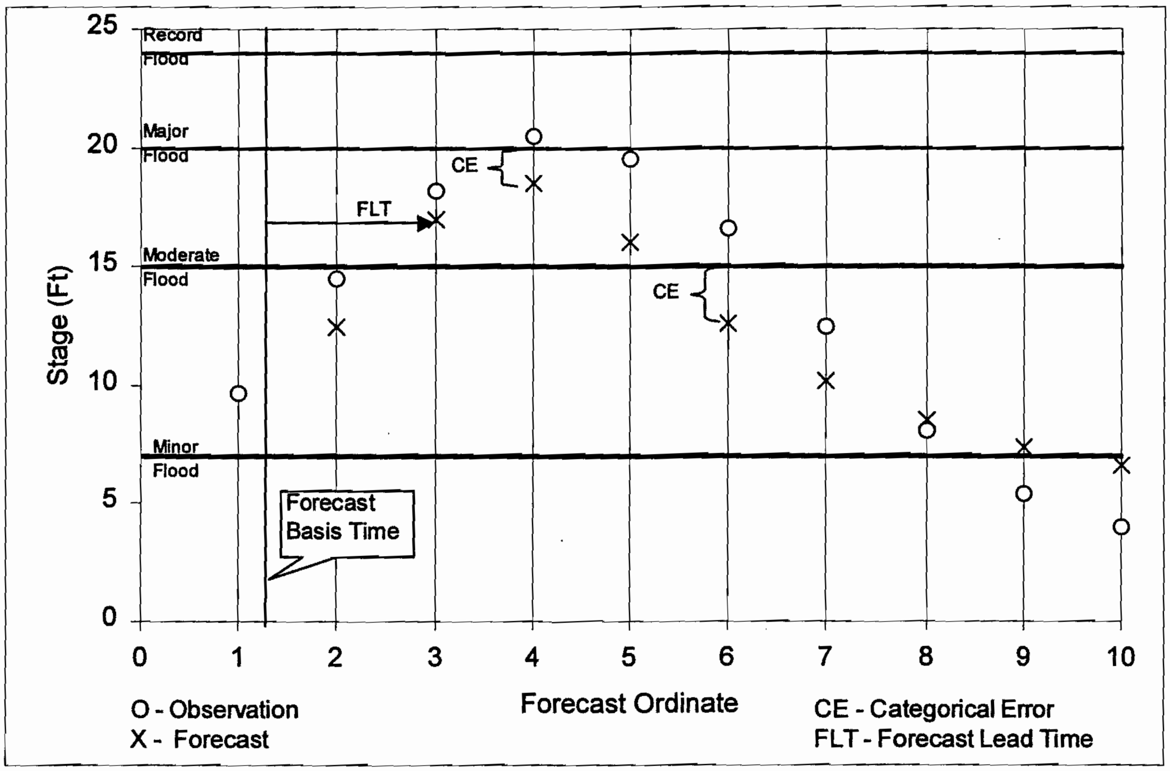


Figure 3 Example of verification calculations for Fcst 1.

Forecast Ordinate	Forecast Category	Observed Category	Categorical Result	Comment
1	No forecast	Minor	Minor No Fcst Miss	Flooding began before forecast issued
2	Minor	Minor	Minor Hit	No Lead Time Computation
3	Moderate	Moderate	Moderate Hit	Calculate Lead Time
4	Moderate	Major	Major Miss	Calculate Cat. Error
5	Moderate	Moderate	Moderate Hit	No Lead Time Computation
6	Minor	Moderate	Moderate Miss	Calculate Cat. Error
7	Minor	Minor	Minor Hit	No Lead Time Computation
8	Minor	Minor	Minor Hit	No Lead Time Computation
9	Minor	No Flood	Minor False Alarm	
10	No Flood	No Flood	No Flood	No Verification

Table 2. Tabulation of verification results for forecast example in Figure 3.

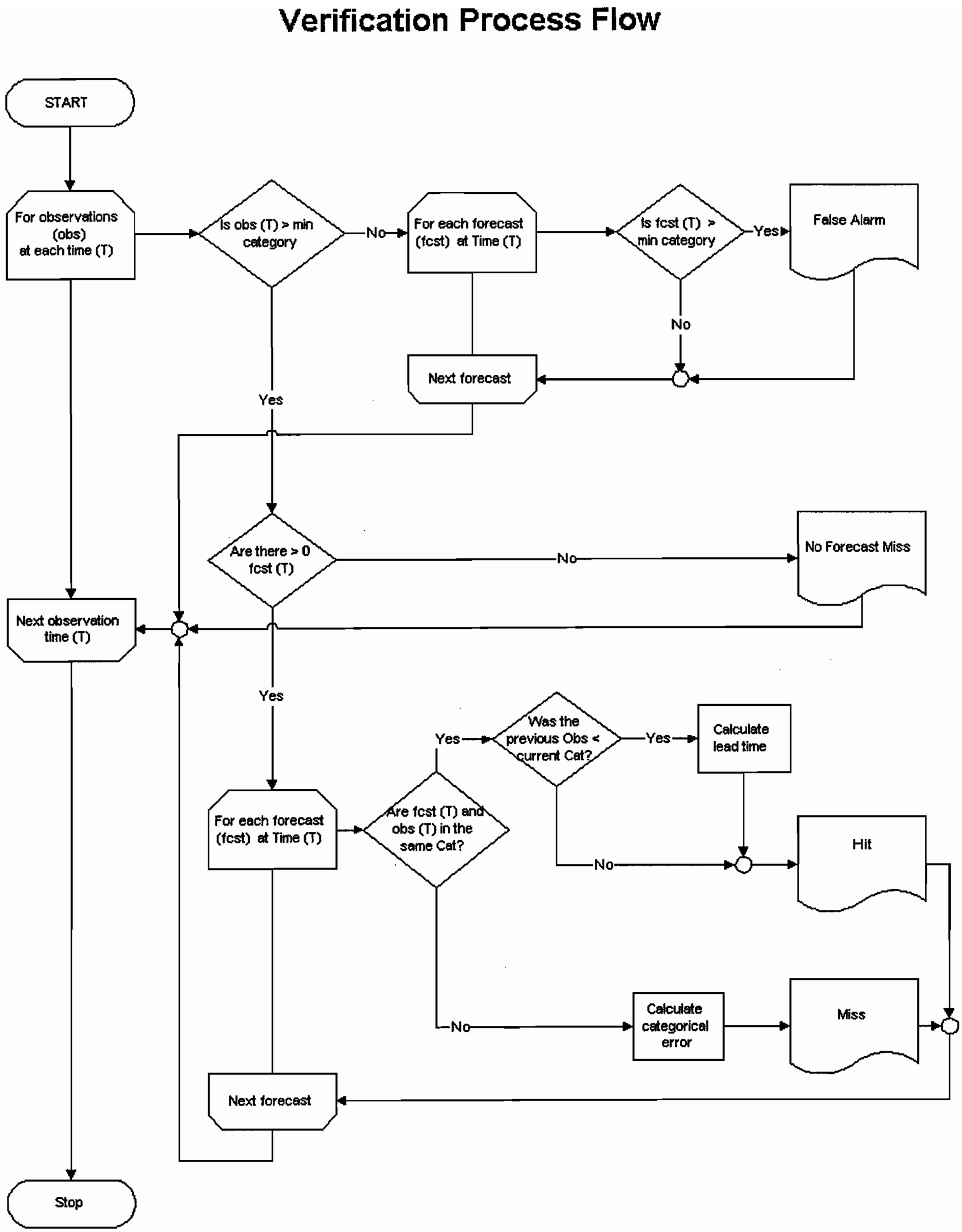

Verification Process Flow
START
For observations (obs) at each time (T)
Is obs (T) > min category
No
For each forecast (fcst) at Time (T)
Is fcst (T) > min category
Yes
False Alarm
No
Next forecast
Yes
Are there > 0 fcst (T)
No
No Forecast Miss
Next observation time (T)
Yes
Yes
Was the previous Obs < current Cat?
Yes
Calculate lead time
For each forecast (fcst) at Time (T)
Are fcst (T) and obs (T) in the same Cat?
No
Hit
No
Calculate categorical error
Miss
Next forecast
Stop

Figure 4 Process flow diagram for categorical verification program

4.3 The Fcstver Program

The forecast verification program, *fcstver*, is written in the C programming language, and utilizes the Xwindows/Motif graphical user interface, Informix embedded SQL (ESQL) for database access, and the gd graphics library, available as shareware on the Web, for creating graphical output files in CompuServe graphics interchange format (GIF). The program is locally configurable using apps-defaults as implemented with other NWS hydrologic software.

Fcstver is set up to provide maximum flexibility in the scope of verification. From a graphical user interface, the user selects a set of forecast points to verify. He can also set the beginning and ending dates for verification. This allows verification of a particular event, or for a selected time period, such as quarterly or annually. Once the user has selected the time period and the forecast points to verify, the program computes statistics and generates two site specific graphics for each forecast point.

The first output graphic, Verification Summary 1, contains two charts (Fig. 5). The top half is a bar chart that presents statistics for each flood category. The bottom half contains an observed stage hydrograph for the entire verification period annotated with all forecast basis times. Also shown on this graphic are the total number of events for each category.

The second output graphic, Verification Summary 2, further breaks down the verification statistics by forecast ordinate (Fig. 6). Verification statistics are tallied by ordinate of the time series forecast, and presented on a separate bar chart for each flood category. This graphic identifies where forecast errors occur with respect to the time the forecast was issued. The naming convention for the summary files are *xxxxx.gif* and *xxxxx2.gif* where xxxxx is the Communications Handbook 5 identifier for the forecast point.

Finally, fcstver generates a text file containing verification values for the entire set of forecast points evaluated. This file is used as input for the executive summary program. The text file is named *yyyy-mm-dd_to_yyyy-mm-dd* where yyyy-mm-dd is the beginning and ending year, month and day of the forecast verification run.

5. Executive Summaries

The forecast verification software creates output which is useful at the RFC level, but is often too detailed for distribution to management. It became obvious that there was an overwhelming need to reduce the verification data to a set of performance metrics that could quickly be interpreted by managers. With this idea in mind, the idea of "Executive Summaries" was employed.

5.1 Executive Summary Methodology

The original executive summary concept consisted of simply aggregating the statistics from all of the forecast points evaluated with a run of the forecast verification program. To this means, a feature was added to fcstver to write out the verification data to a text output file. An executive summary program, *execfcstver*, was then written to read this text file and create executive summary graphics in GIF format. This yielded two graphic products similar to the individual forecast point products, but contained cumulative statistics for all forecast points verified. These products are labeled as "Executive Summary 1" and "Executive Summary 2" (Figs. 7 and 8).

While this verification aggregation condensed the data considerably, and seemed appropriate for presentation to management, we felt it was necessary to further reduce the data to a few easily understandable performance metrics. At this point we researched the standard statistics used in weather verification - Probability of Detection (POD), False Alarm Ratio (FAR)and Lead Time.

The NWS (Halmstad,1996) offers the following definitions:

> The POD is the number of warned severe local storm events divided by the total number of severe local storm events reported.
>
> The FAR is the number of unverified county warnings divided by the total number of county warnings issued.

Applying these definitions to our categorical flood forecast verification scheme, we derived the following equations:

(1) POD = Hits / (Hits + Misses + No Forecast Misses)

(2) FAR = False Alarms / (False Alarms + Hits)

The categorical lead time (CLT) is still defined as the time from issuance of a forecast ordinate to the time of the forecast ordinate for qualifying hits.

(3) CLT = Valid Time - Basis Time

Using the applied definitions for POD and FAR, we arrived at two additional executive summary graphics. "Executive Summary 3" shows POD and FAR by category (Fig. 9). These statistics stand true to the previously defined values of hits, misses and false alarms. Since one must forecast within the observed category to be credited with a hit, there is no partial credit given for forecasts that correctly identify a flood but are in the wrong category. From this information one can easily show success of predicting floods of different magnitudes (or categories). The number of events for each category are also displayed on this graphic.

The final graphic product, "Executive Summary 4" represents forecast performance in its simplest terms (Fig. 10). This product shows POD and FAR based on a flood / no flood evaluation. A hit is redefined to be an ordinate where the observation is at minor flood or above, and the forecast is also at minor flood or above. Other categories are ignored. Likewise, misses and false alarms are evaluated based on above or below minor flood criteria. The resulting product shows a single value for POD and FAR.

The number of ordinates in each forecast issuance varies by RFC as well as forecast point location, depending on river response and local office policies. Some forecasts extend five days into the future, while others are only issued through day three or four. Forecast skill generally decreases with time, so consequently, forecast error increases. The final verification statistics considered all forecasts, so in most cases a location where five-day forecasts are issued will not compare favorably to locations where forecasts only go out three days. To even this playing field, the final verification numbers were stratified by days, and new sets of "Executive Summary 3 and 4" graphics were created for each forecast day. These graphics are identical in appearance to the original Executive Summaries 3 and 4.

The reduction of verification statics into commonly used performance measures for weather events has many benefits. NWS management can easily relate to terms such as FAR and POD, but these numbers are not meant to be, and never should be compared to FAR and POD of weather events.

6. Conclusion

It is a goal of the National Weather Service to improve the accuracy and timeliness of river forecasts. We constantly seek to realize these improvements through the infusion of technology and science into our forecast systems. It is, however, difficult to evaluate the impact of these improvements on the accuracy of our forecasts without meaningful performance metrics. A categorical based verification scheme such as the CFFVS, can provide valuable insight into the RFC forecast service. This evaluation is not only valid at a single site, but is also applicable when accumulated for multiple sites.

Acknowledgments

The authors wish to acknowledge Dr. David G. Morris for originating the concept of categorical river forecast verification, and for providing the impetus for implementing this verification scheme in the Southern Region. We also want to acknowledge the additional members of the Southern Region RFC verification team - Billy Olsen, John Schmidt, Brad Gimmestad, Bob Stucky, and Ben Weiger for their diligent work in expounding on the original concept, and arriving at the final verification metrics.

References

Halmstad, J.T., 1996: Severe local storm warning verification for 1995. *NOAA Technical Memorandum NWS-SPC-1.*

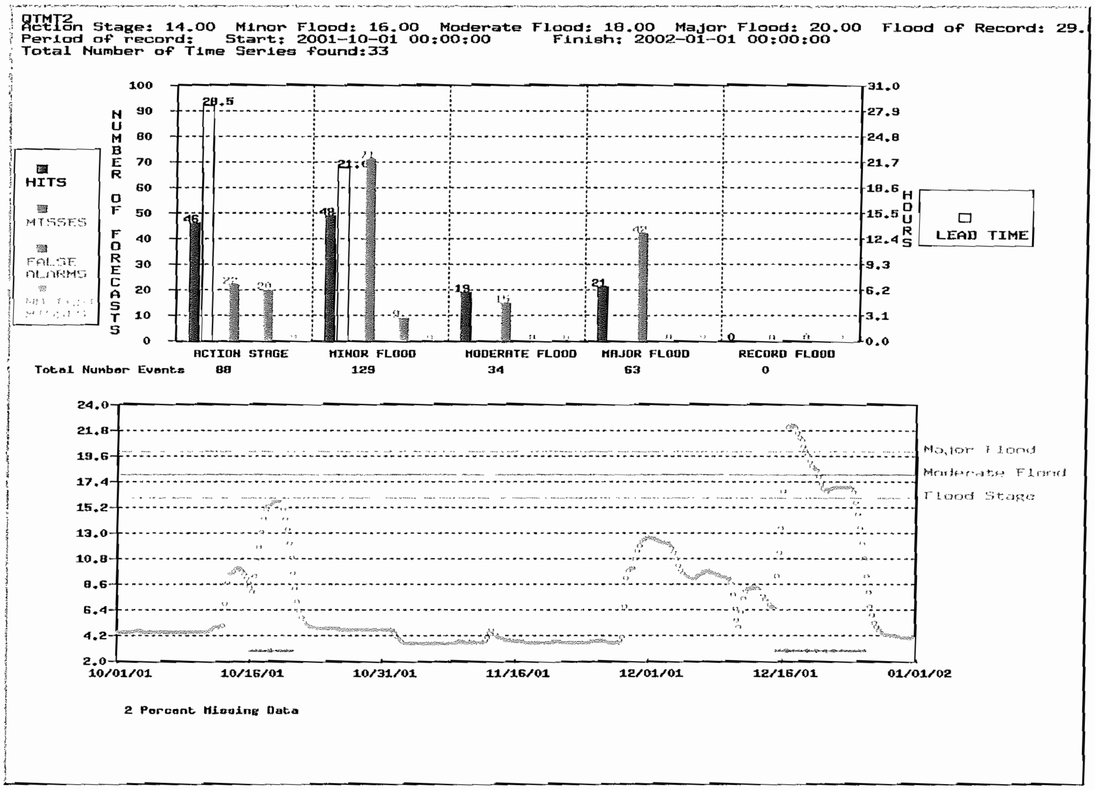
Morris, D.G., 1988: A Categorical, Event Oriented, Flood Forecast System for National Weather Service Hydrology. *NOAA Technical Memorandum NWS HYDRO-43.*

National Weather Service OS/HSD, RFC Verification Implementation Plan, Jan 2001.

National Weather Service Office of Hydrology, Weather Service Hydrology Handbook No. 1. Version 1.3, March 1998.

Reed, W. B., B. G. Olsen, and J. A. Schmidt, 1997: An Evaluation of River Forecast Model Output for Simulations With and Without Quantitative Precipitation Forecasts (QPF). *NOAA Technical Memorandum NWS SR-187.*

Wasko, R., W. E. Lawrence, and B. G. Olsen, 2001: River Forecast Verification at the ABRFC. *NOAA Technical Memorandum NWS SR-212.*



QTMT2
Action Stage: 14.00 Minor Flood: 16.00 Moderate Flood: 18.00 Major Flood: 20.00 Flood of Record: 29.
Period of record: Start: 2001-10-01 00:00:00 Finish: 2002-01-01 00:00:00
Total Number of Time Series found:33

Figure 5 Verification Summary 1

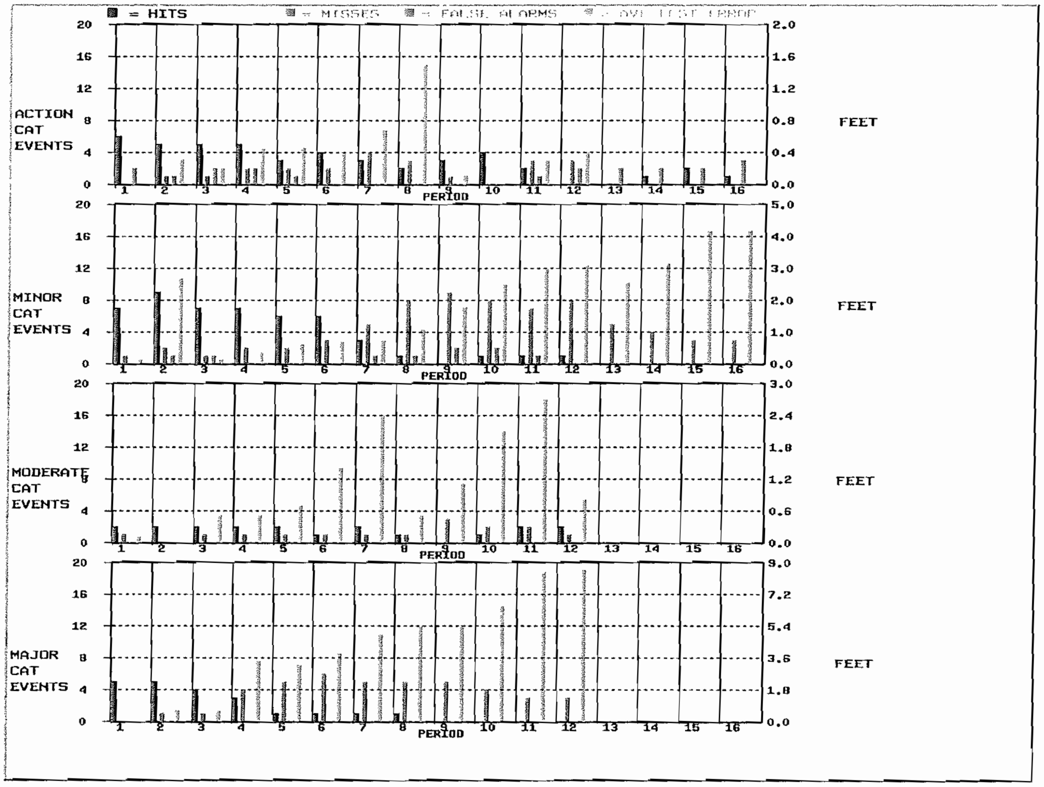


Figure 6 Verification Summary 2

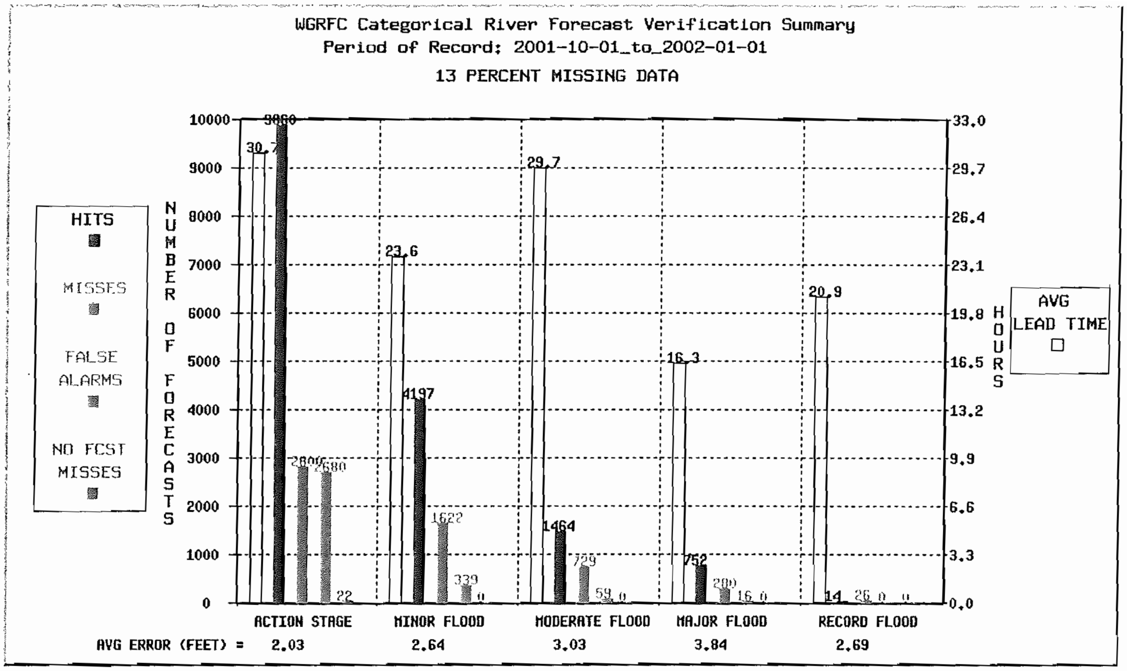


Figure 7 Executive Summary 1.

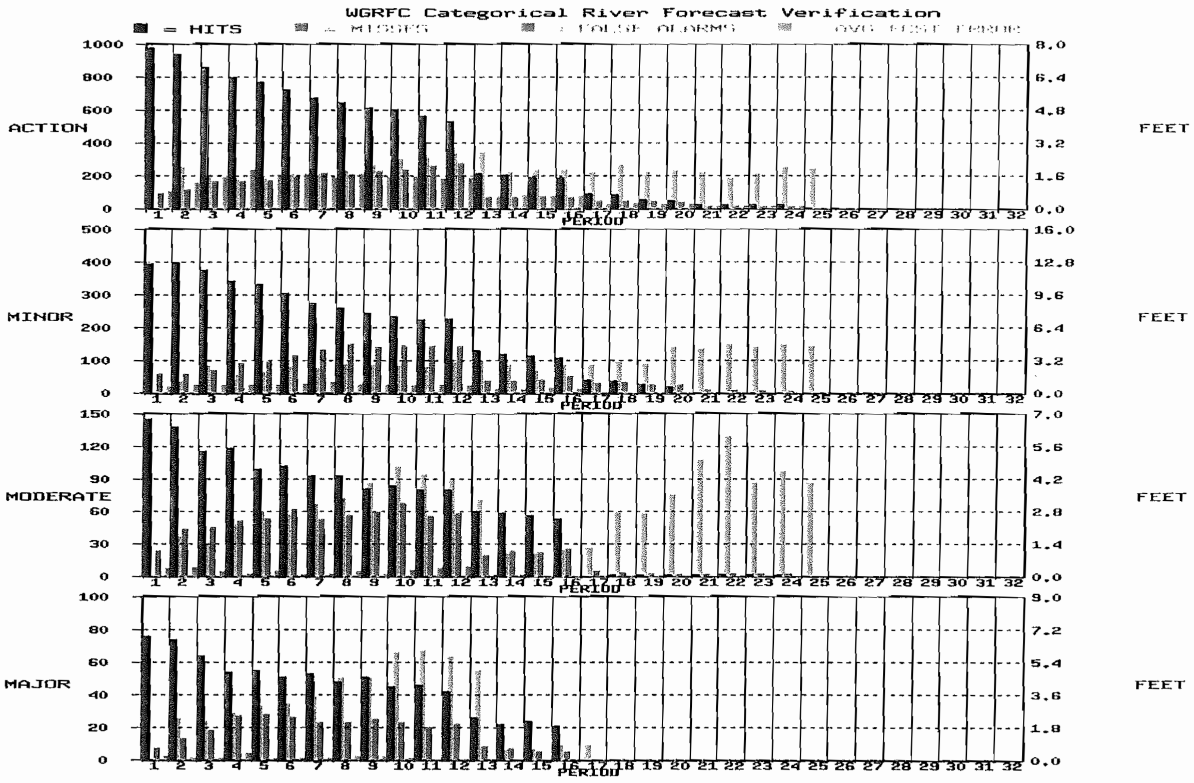


Figure 8 Executive Summary 2.

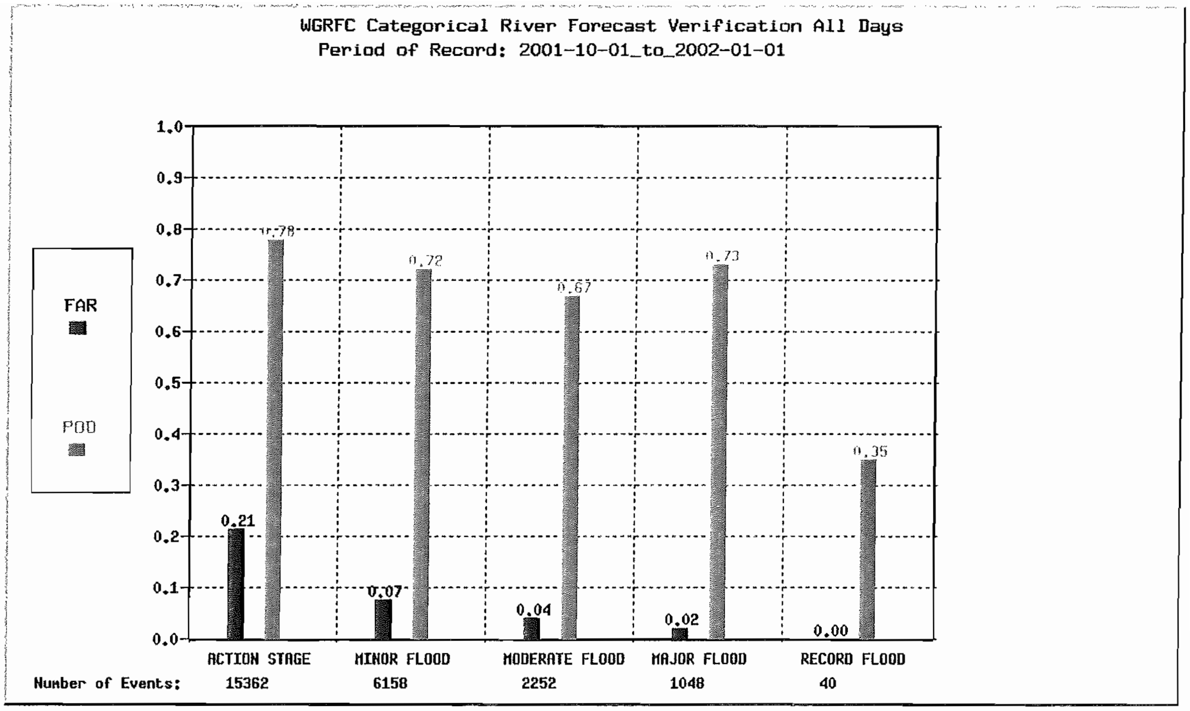


Figure 9 Executive Summary 3.

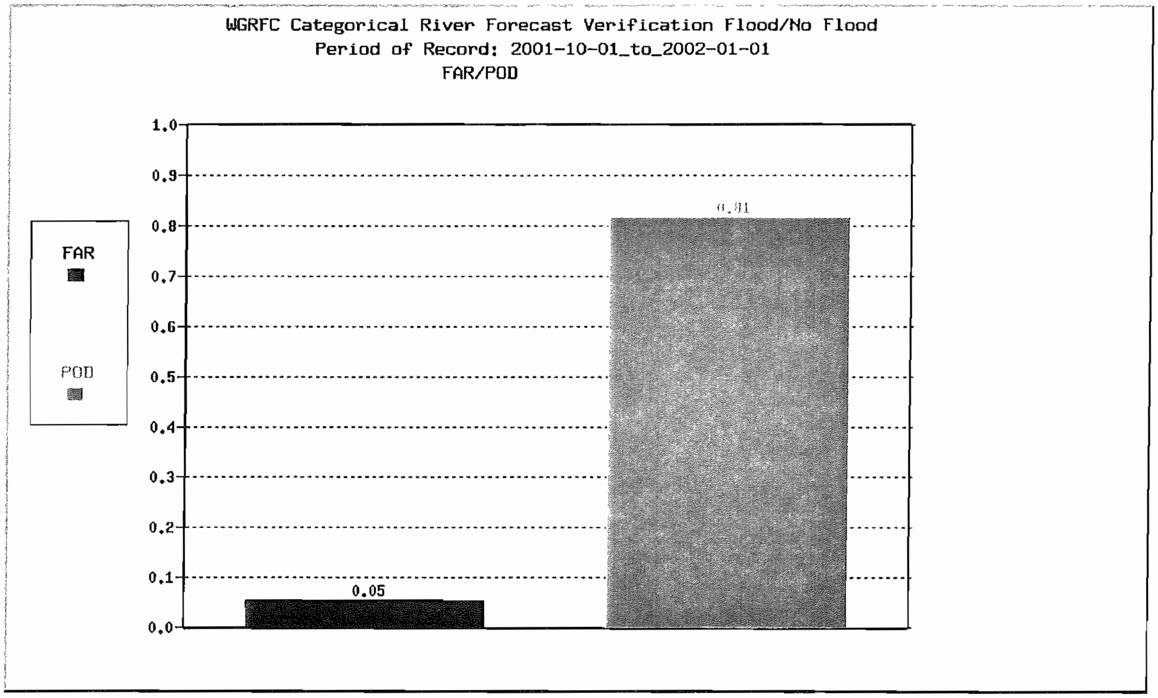


Figure 10 Executive Summary 4.

Appendix A

SAMPLE XSETS GENERATED SHEF ENCODED RIVER FORECAST

```
FTWRVFSHV DEF
TTAA00 KFWR DDHHMM
RIVER FORECASTS
NATIONAL WEATHER SERVICE...WEST GULF RFC...FORT WORTH, TEXAS
1036 AM CST TUE DEC 26 2000
:
:      SABI
:
:
:MINEOLA
:
:LATEST STAGE    11.77 FT AT 715 AM CST ON 1226
.ER MLAT2    1226 C DC200012261036/DH12/HGIFF/DIH6
:RIVER FORECAST      6AM       NOON        6PM       MDNT
.E1 :1226:              /      12.5/      13.8/      14.5
.E2 :1227:   /      14.9/      15.1/      15.5/      15.7
.E3 :1228:   /      15.5/      15.4/      15.0/      14.5
.E4 :1229:   /      13.8
:  MLAT2 (AS 12 FS 14)
:
:QUITMAN 2E
:
:LATEST STAGE     3.63 FT AT 700 AM CST ON 1226
.ER QTMT2    1226 C DC200012261036/DH12/HGIFF/DIH6
:RIVER FORECAST      6AM       NOON        6PM       MDNT
.E1 :1226:              /       5.1/       9.8/      14.4
.E2 :1227:   /      15.5/      15.7/      15.8/      16.2
.E3 :1228:   /      15.4/      14.8/      14.2/      14.0
.E4 :1229:   /      13.7
:  QTMT2 (AS 14 FS 16)
:
: apg
:
:...END OF MESSAGE...
NNNN
```

Appendix B

SCHEMA FOR ARCHIVE DATABASE FCSTHEIGHT AND HEIGHT TABLES

Archive database: schema of height table:

```
lid char(8)
pe char(2)
dur integer
ts char(2)
extremum char(1)
obstime datetime year to second
value float
shef_qual_code char(1)
quality_code integer
man_edited char(1)
processed_code integer
product_id char(10)
producttime datetime year to second
postingtime datetime year to second
```

Archive database: schema of fcstheight table

```
lid char(8)
pe char(2)
dur integer
ts char(2)
extremum char(1)
probability float
validtime datetime year to second
basistime datetime year to second
value float
shef_qual_code char(1)
quality_code integer
man_edited char(1)
processed_code integer
product_id char(10)
producttime datetime year to second
postingtime datetime year to second
```